Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/080651

An Efficient Gene Selection Technique based on Fuzzy C-means and Neighborhood Rough Set

Jiucheng Xu[1,2], *Tianhe Xu*[1,*], *Lin Sun*[1,2] *and Jinyu Ren*[1]

[1] College of Computer and Information Engineering, Henan Normal University, 453007 Xinxiang, China
[2] Engineering Technology Research Center for Computing Intelligence and Data Mining, Henan Province, China

Received: 21 Nov. 2013, Revised: 19 Feb. 2014, Accepted: 20 Feb. 2014
Published online: 1 Nov. 2014

Abstract: Selecting genes from microarray gene expression datasets has become an important research, because such data typically consist of a large number of genes and a small number of samples. Avoiding information loss, neighborhood mutual information is used to evaluate the relevance between genes in this work. Firstly, an improved Relief feature selection algorithm is proposed to create candidate feature subsets. Then, the cohesion degree of the neighborhood of an object and coupling degree between neighborhoods of objects are defined based on neighborhood mutual information. Furthermore, a new initialization method of cluster centers for the Fuzzy C-means (FCM) algorithm is proposed. FCM is a method that allows one piece of data to belong to two or more clusters. Moreover, in view of neighborhood rough set is an effective tool to extract and select features, a novel algorithm for gene selection based on FCM algorithm and neighborhood rough set is proposed. Finally, to evaluate the performance of the proposed approach, we apply it to five well-known gene expression datasets. Experimental results show that the proposed approach can select genes effectively, and can obtain high and stable classification performance.

Keywords: Fuzzy C-means, neighborhood rough set, neighborhood mutual information, Relief algorithm, gene selection

1 Introduction

Microarray technology has made it possible to simultaneously measure the expression levels of large numbers of genes in a short time [1,2]. However, among the large amount of genes presented in microarray gene expression datasets, only a small fraction of them is effective for performing a certain diagnostic test. So, the curse of dimensionality caused by high dimensionality and small sample size of tumor dataset seriously challenges the tumor classification [3,4]. How to select important gene subsets from thousands of genes in gene expression profiles dataset to drastically reduce the dimensionality of tumor dataset is the key step to address this problem.

Many gene selection methods have been proposed for the analysis of gene expression datasets [5,6]. Usually, the feature selection methods can be divided into three broad categories: filter, wrapper, and embedded methods [7,8,9]. The filter method is to design a measure independent of a specific classification algorithm. Thus features that accurately present the original data set can be identified. The filter methods include correlation-based feature selection [10], t-test, information gain, mutual information, and entropy-based methods [11]. However, they ignore feature dependencies, resulting in poor classification performance. Wrapper methods focus on improving classification accuracy of pattern recognition problems and typically perform better than filter methods. However, wrapper methods are more time-consuming than filter methods [12]. Embedded techniques combine filter methods and wrapper methods. The advantage of the embedded algorithms is that they take the interactions with the classifiers into account.

Clustering analysis is an important technique in pattern recognition, which aims to divide a data set into several clusters [13]. The clustering algorithms can be broadly classified as Hard, Fuzzy, Possibilistic, and Probabilistic [14]. The ability of clustering methods is to extract groups of genes with similar functions from huge datasets according to the fact that genes with similar functions evince similar expression patterns of co-regulation [15,16]. Intuitively, genes in a cluster are more correlated with each other, whereas genes in

[*] Corresponding author e-mail: tianhe1107@163.com

different clusters are less interdependent [17]. K-means is one of the most popular hard clustering algorithms which partitions data objects into k clusters where the number of clusters, k is decided in advance according to application purposes. However, hard clustering methods which assign each gene exactly to one cluster are poorly suited to the analysis of gene expression datasets because in such datasets the clusters of genes frequently overlap [4]. To overcome the limitations of these hard clustering methods, fuzzy clustering has been widely studied, in which a data point is associated with multiple clusters to different extents based on its membership values to these clusters [18, 19]. FCM algorithm is one of the most popular fuzzy clustering techniques because it is efficient, straightforward, and easy to implement.

In FCM, the objective is to minimize the sum of cluster variations, which depends on the distances between data and the cluster centers [19, 20]. Since the Euclidean distance is used, the cluster structures are all hyper spherical. In order to improve the ability to detect cluster structures of other shapes, many researchers extended FCM by redefining the distance and cluster centers [19, 21, 22]. FCM clustering is an effective algorithm, but the random selection in cluster centers makes iterative process falling into the local optimal solution easily. The algorithm with random initialization method needs to be rerun many times with different initializations in an attempt to find a good solution. Furthermore, random initialization method works well only when the number of clusters is small and chances well that at least one random initialization is close to a good solution [23]. Therefore, how to choose proper initial cluster centers is extremely important as they have a direct impact on the formation of final clusters. In this paper, an initialization method of cluster centers for the FCM algorithm is proposed which based on the cohesion degree of neighborhood of an object and the coupling degree between neighborhoods of objects.

Rough set (RS) theory, proposed by Pawlak [24], can be seen as a new mathematical approach for vague questions. It has been successfully applied to pattern recognition, expert system, machine learning, knowledge discovery, decision analysis and data mining. RS has been applied mainly in mining tasks like classification, clustering and feature selection [25, 26]. The gene expression datasets often consist of small number of samples and large number of genes. The curse of dimensionality makes it necessary to reduce the computation cost and improve the classification accuracy. RS provides a feasible way to deal with redundancy to find out a minimum set of relevant attributes that describe the dataset as well as all the original attributes do [25]. However, the feature reduction in classical RS must discretize the attributes before reduction, which maybe leads to information loss. As we know that gene expression data is numerical, in order to deal well with the datasets, we have to avoid the discretization. Hu et al. [27] introduced the neighborhood rough set (NRS) model, which based on classical RS and neighborhood relevance, avoiding the discretization procedure, so no information loss occurs. Zhang et al. [25] designed a tumor classification method based on wavelet packet transforms and neighborhood rough set. Wang et al. [28] proposed a novel feature selection method by combining PNN classifier ensemble with neighborhood rough set. A quick search of biological literatures shows that NRS is still seldom used in bioinformatics. Based on FCM and NRS, a novel gene selection method is proposed in this paper. Firstly, an improved Relief feature selection algorithm which based on neighborhood mutual information is proposed to sequence genes, and generate candidate feature subsets. Then an initialization method of cluster centers for the FCM algorithm is proposed which based on the cohesion degree of neighborhood of an object and the coupling degree between neighborhoods of objects. Moreover, the significance of attributes based on neighborhood rough set is defined. Finally, a novel gene selection algorithm (NMINR-FCM) is proposed, which based on FCM and neighborhood rough set, to obtain both better performance and superior classification accuracy.

The structure of the rest of this paper is as follows: Section 2 introduces the concepts of FCM clustering and neighborhood rough set. An effective and efficient gene selection method NMINR-FCM is proposed in Section 3. To evaluate the performance of the proposed algorithm, we apply it to five gene expression datasets. The experimental results are presented in Section 4. Finally, the conclusion is drawn in Section 5.

2 Related work

2.1 Fuzzy C-means Clustering

FCM clustering algorithm is developed by Dunn [29] and later refined by Bezdek [30], is an unsupervised fuzzy clustering algorithm with multiple applications, ranging from attribute analysis, to clustering and classifier design. Let the sample set be $X = \{x_1, x_2, \ldots, x_n\}$ where n is the number of sample. FCM algorithm divide the sample set X into c $(2 \leqslant c \leqslant n)$ classes is $U = [u_{ij}]_{c\times n}$, where $u_{ij}(1 \leqslant i \leqslant c, 1 \leqslant j \leqslant n)$ is the fuzzy membership degree of the jth sample x_j belongs to the ith class, and u_{ij} should satisfy the following constraint

$$\sum_{i=1}^{c} u_{ij} = 1, 0 \leqslant u_{ij} \leqslant 1, 1 \leqslant i \leqslant c, 1 \leqslant j \leqslant n. \quad (1)$$

The objective function of FCM algorithm is defined as

$$minJ_m(U,P) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m d_{ij}^2, \quad (2)$$

where $d_{ij} = ||x_j - p_i||$ is the distance between x_j and p_i, p_i is the center of the ith class. The fuzziness of the membership is controlled by m which takes value higher than 1. The closer is the m value to 1, the more crisper the membership values are. As the values of m become progressively higher, the resulting memberships become fuzzier [31]. Pal and Bezdek advised that m should take value between 1.5 and 2.5 [32]. FCM algorithm is an process to minimize the objective function min $J_m(U,W)$. To achieve $minJ_m(U,W)$ should meet the following conditions

$$u_{ij} = (\sum_{k=1}^{c}(\frac{d_{ij}}{d_{kj}})^{2/(m-1)})^{-1}, \tag{3}$$

$$p_i = \frac{\sum\limits_{j=1}^{n} u_{ij}^m x_j}{\sum\limits_{j=1}^{n} u_{ij}^m}. \tag{4}$$

A iteration alternating between equation (3) and (4) adjust u_{ij} and p_i until the change in J_m falls below a threshold ε or a maximal number of iterations t is reached. In this study, we chose ε = 0.001 and t = 100.

2.2 Neighborhood rough set

In order to effectively cope with continuous attributes, avoiding the information loss which caused by discretization, Hu et al. [33] proposed neighborhood rough set based on classical rough sets and the concept of neighborhood. In this subsection, we will introduce the basic concepts of neighborhood rough set. The basic concepts of neighborhood rough set are explained as follows.

Given arbitrary $x_i \in U$ and $B \subseteq A$, $\delta \geqslant 0$ is a constant, then the neighborhood of sample x_i is denoted by

$$\delta_B(x_i) = \{x \in U | \Delta_B(x,x_i) \leq \delta\}, \tag{5}$$

where Δ is a distance function on U.

Let $A = \{a_1, a_2, \ldots, a_n\}$ be a discrete random variable. $P(a_i)$ is the probability of a_i, the entropy of A is denoted by

$$H(A) = -\sum_{i=1}^{n} p(a_i)\log p(a_i). \tag{6}$$

Let $U = \{x_1, x_2, \ldots, x_n\}$ be a set of samples described with gene set F, and $x_i \in R^N$, $S \subseteq F$ is a subset of genes, the neighborhood of sample x_i in S is denoted by $\delta_S(x_i)$. The neighborhood uncertainty of x_i is denoted by

$$NH_\delta^{x_i}(S) = -\log\frac{||\delta_S(x_i)||}{n}, \tag{7}$$

and the average uncertainty of the set of samples is computed as

$$NH_\delta(S) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{||\delta_S(x_i)||}{n}. \tag{8}$$

If $\delta = 0$, then $NH_\delta(S) = H(S)$, where $H(S)$ is Shannon entropy.

Given $R,S \subseteq F$ are two subsets of genes, the neighborhood of sample x_i in gene subspace $S \cup R$ is denoted as $\delta_{S\cup R}(x_i)$, then the joint neighborhood entropy of $S \cup R$ is computed as

$$NH_\delta(R,S) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{||\delta_{S\cup R}(x_i)||}{n}. \tag{9}$$

The conditional neighborhood entropy of R is defined as

$$NH_\delta(R|S) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{||\delta_{S\cup R}(x_i)||}{||\delta_S(x_i)||}. \tag{10}$$

Hence the following property holds, $NH_\delta(R|S) = NH_\delta(R,S) - NH_\delta(S)$.

3 Efficient gene selection algorithm

3.1 Neighborhood Mutual Information Measure

Generally speaking, Euclidean distance, Pearson's correlation coefficient and mutual information are widely used as the measure to compute relevance between attributes. However, for measuring the correlation between genes, Euclidean distance is not effective enough to describe functional similarity such as positive or negative correlation in values [33]. Thus, Pearson's correlation coefficient [34] is put forward by some researchers. Empirical studies have shown that it may assign a high similarity score to a pair of dissimilarity genes. There is a problem to employ mutual information in gene evaluation due to the difficulty in estimating probability density of genes. However, most methods are not able to effectively cope with continuous attributes, which is also a distinctive characteristic of gene expression datasets. When applied to the continuous attributes, conventional methods commonly discretize the continuous data into a finite number of intervals for data mining. But discretization may lead to information loss [27]. Hu et al. [33] proposed neighborhood mutual information to cope with continuous attributes, evaluate the relevance between attributes. The neighborhood mutual information combines the concept of neighborhood with information theory, and generalizes Shannon's entropy to numerical information.

Let $R,S \subseteq F$ are two subsets of genes, then the neighborhood mutual information of R and S is denoted by

$$NMI_\delta(R;S) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{||\delta_R(x_i)||\cdot||\delta_s(x_i)||}{n||\delta_{S\cup R}(x_i)||}. \tag{11}$$

The following properties hold
(1) $NMI_\delta(R;S) = NMI_\delta(S;R)$;
(2) $NMI_\delta(R;S) = NH_\delta(R) + NH_\delta(S) - NH_\delta(R,S)$;
(3) $NMI_\delta(R;S) = NH_\delta(R) - NH_\delta(R|S) = NH_\delta(S) - NH_\delta(S|R)$.

3.2 The Improved Relief Algorithm

Relief as a kind of attribute ordering algorithm has been widely applied in the field of feature selection. Its core idea is to distinguish similar samples as the standard of evaluation attribute importance, and thus gives the attribute weights in classification. The advantage of this algorithm is less computational complexity, considering the correlation between attributes to a certain extent. For arbitrary sample, searching out two class neighbors which nearest to this sample, one kind with the same classes of groups (called nearest hit), and another kind is the category with its distinct groups (called nearest miss). Then the search process in a sample of nearest neighbors is to take the distance between the two samples as the standard. In the Relief algorithm, all the attributes are involved in the distance calculation process. However, in gene expression datasets, only a small number of genes associated with the sample type, the vast majority of genes as noise properties exist. If use Relief algorithm to select the gene expression datasets directly, will make the noise drowned out the useful information, resulting in the classification weights calculated of genes deviate from the true value. The RFE_Relief algorithm presented in [35], firstly, computed the attribute classification weights using Relief algorithm, and then remove the attribute with the minimum weight, and so on, the effect of noise properties reducing gradually. However, this algorithm did not take into account the relationship between features of each sample, which affects the accuracy of classification. In this paper, we improve the RFE_Relief algorithm and propose an improved Relief algorithm to select sample classification genes, which uses the neighborhood mutual information to measure the correlation between genes. The algorithm is described as follows:

Algorithm 1. NRFE_Relief algorithm

Input: Sample set $X = \{x_1, x_2, \ldots, x_M\}$ and gene set $G = \{g_1, g_2, \ldots, g_N\}$

Output: Gene subset B

Step1: Set the weight vector W;

Step2: For arbitrary sample x_i (i = 1, 2, ..., M, M as sample number), search its P nearest hit and P nearest miss;

Step3: For any gene g_j (j = 1, 2, ..., $|G|$, $|G|$ as gene number), calculate the weight of it: $W(g) = W(g) - diff(g,x_i,H)/P + diff(g,x_i,M)/P$, where H is the nearest neighbors with the same type with sample x_i, M is the nearest neighbor has different categories with sample x_i. The function $diff(g,x_i,x_j)$ is used to calculate the difference between samples x_i and x_j of gene g, $diff(g,x_i,x_j) = |NMI_\delta(g;x_i) - NMI_\delta(g;x_j)|$;

Step4: Finding the locate of attribute with minimum weight, according to the $c = argminW$;

Step5: $B = G - \{g_c\}$;

Step6: END.

In actual operation process, the 10% of the total attributes was removed to speed up the operation of the algorithm, and neighbor number K = 15.

3.3 Cohesion Degree of the Neighborhood of An Object and Coupling Degree between Neighborhoods of Objects

Formally, the structural data used for classification learning can be written as an information system, denoted by $IS = < U,A,V,f >$, where U is the is the nonempty set of samples $\{x_1, x_2, \ldots, x_M\}$, called a universe; A is a set of attributes $\{a_1, a_2, \ldots, a_n\}$ to characterize the samples; V is the union of all attribute domains, ie., $V = \cup V_a$, where V_a is the value domain of attribute a and $V \subset R$; f is a mapping called an information function such that for any $x \in U$ and $a \in A$, $f(x,a) \in V_a$.

Definition 3.1. Let $IS = < U,A,V,f >$ be be a numeric information system and $B \subseteq A$. For any $x_i, x_j \in U$, the correlation between this two attributes is denoted by $NMI_\delta(x_i,x_j)$, the average correlation among attributes is defined as

$$\bar{x} = \frac{2}{|U|(|U|-1)} \sum_{i=1}^{|U|-1} \sum_{j=i+1}^{|U|} NMI_\delta(x_i,x_j), \qquad (12)$$

the size of $\bar{x}$ measures the distribution of objects in U. The greater $\bar{x}$ is, the looser distribution among objects is. Hence, in the rest of this paper, we use $\bar{x}$ to denote the size of neighborhood of objects, that is $\varepsilon = \bar{x}$.

Definition 3.2. Let $IS = < U,A,V,f >$ be be a numeric information system, $B \subseteq A$ and $X \in U$, the lower and upper approximations of X in U with respect to B are defined as

$$\underline{B}X = \{x_i | \delta_B(x_i) \subseteq X, x_i \in U\}, \qquad (13)$$

and

$$\overline{B}X = \{x_i | \delta_B(x_i) \cap X \neq \varnothing, x_i \in U\}, \qquad (14)$$

$\underline{B}X$ is a set of objects whose neighborhood belongs to X with certainty, while $\overline{B}X$ is a set of objects whose neighborhood possibly belongs to X.

Obviously, $\underline{B}X \subseteq X \subseteq \overline{B}X$. The boundary region of X in the approximation space is defined as

$$BNX = \overline{B}X - \underline{B}X. \qquad (15)$$

Definition 3.3. Let $IS = < U,A,V,f >$ be be a numeric information system, $B \subseteq A$. For any $x_i \in U$, the cohesion degree of $\delta_B(x_i)$ is defined as

$$Cohesion(\delta_B(x_i)) = \frac{|\underline{B}(\delta_B(x_i))|}{|\overline{B}(\delta_B(x_i))|}, \qquad (16)$$

where $0 < Cohesion(\delta_B(x_i)) \leq 1$.

The greater $Cohesion(\delta_B(x))$ is, the less boundary region of neighborhood of object x is, which means that x is a better cluster center of its neighborhood. Therefore, x is likely taken as an initial cluster center in U.

Definition 3.4. Let $IS = <U, A, V, f>$ be be a numeric information system, $B \subseteq A$. For any $x_i, x_j \in U$, the coupling degree of $\delta_B(x_i)$ and $\delta_B(x_j)$ is defined as

$$coupling(\delta_B(x_i), \delta_B(x_j)) = \frac{|\delta_B(x_i) \cap \delta_B(x_j)|}{|\delta_B(x_i) \cup \delta_B(x_j)|}, \quad (17)$$

where $0 < coupling(\delta_B(x_i), \delta_B(x_j)) \leq 1$.

The greater $coupling(\delta_B(x_i), \delta_B(x_j))$ is, the more possibly x_i and x_j belong to the same cluster will be. In this paper, if $(\delta_B(x_i), \delta_B(x_j)) > \varepsilon$, we consider that x_i and x_j belong to the same cluster. On the contrary, x_i and x_j are likely taken as initial cluster centers.

The cohesion degree and coupling degree reflect the intracluster similarity and the intercluster similarity, respectively. In this section, based on the cohesion degree of neighborhood of an object and the coupling degree between neighborhoods of objects, an initialization method of cluster centers for the FCM algorithm is described as Algorithm 2:

Algorithm 2. An initialization method of cluster centers for the FCM algorithm

Input: $S = <U, A, V, f>$ and K

Output: Cluster Centers

Step1: Initialize $Centers = \varnothing$ and $Tempcohesion = \varnothing$;

Step2: Compute ε;

Step3: For any $x \in U$, compute $Cohesion(\delta_B(x_i))$, $Centers = Centers \cup \{x\}$ and $Tempcohesion = Tempcohesion \cup \{x\}$, where x satisfies $Cohesion(\delta_B(x_i))$ $= max_{i=1}^{|U|}\{Cohesion(\delta_B(x_i))\}$, the first initial cluster center is selected;

Step4: Find the next most coherent object x, where x satisfies $Cohesion(\delta_B(x_i))$ $= max\{Cohesion(\delta_B(x_i))|x_i \in U - Tempcohesion\}$;

Step5: For any $x' \in Centers$, if $coupling(\delta_B(x'), \delta_B(x)) < \varepsilon$, then $Centers = Centers \cup \{x\}$ and $Tempcohesion = Tempcohesion \cup \{x\}$;

Step6: If $|Centers| < K$, then goto step4, else goto step7;

Step7: END.

3.4 The Significance of Attributes based on Neighborhood Mutual Information

The significance of attributes can be used as heuristic information in greedy algorithm to compute a minimal attribute reduct. In this paper, the significance of an attribute is proposed based on neighborhood mutual information. If the neighborhood mutual information is larger, the two attribute sets are closely related. If the neighborhood mutual information becomes zero, the two attributes are independent.

Definition 3.5. Let A_i and A_j be two attributes, $i, j \in \{1, 2, \ldots, p\}$, $i \neq j$, the significant factor of an attribute A within an attribute cluster $C = \{A_j | j = 1, 2, \ldots, p\}$ is defined as

$$F(A_i) = \sum_{i=1}^{p} NMI_\delta(A_i; A_j), \quad (18)$$

where $NMI_\delta(A_i; A_j)$ is the correlation between A_i and A_j. Based on the concept of $F(A_i)$, we introduce the concept of the *core*, which is an attribute with the highest significant factor in an attribute cluster. The *core* of an attribute cluster $C = \{A_j | j = 1, 2, \ldots, p\}$, denoted by $\eta(C)$, which is an attribute, say A_i, in that cluster such that $F(A_i) \geq F(A_j)$, for all $j \in \{1, 2, \ldots, p\}$.

3.5 The Description of the Improved Gene Selection Algorithm

In this study, a novel gene selection algorithm based on FCM algorithm and neighborhood rough set (NMINR-FCM) is proposed. Firstly, an improved Relief feature selection algorithm is proposed to sequence genes, and generate candidate feature subsets. Then, utilizing FCM algorithm which the cluster centers are initialized based on Algorithm 1 to cluster candidate feature gene subsets. Furthermore, the relevancies between attributes are evaluated by neighborhood mutual information, and the significance of attributes is defined. Finally, select the attribute to represents the cluster which has the highest significance within each cluster. The detailed processing steps in the proposed algorithm are illustrated in flow chart form in Fig.1 and can be described as follows:

Algorithm 3. An efficient gene selection based on FCM algorithm and neighborhood rough set (NMINR-FCM)

Input: Sample set $X = \{x_1, x_2, \ldots, x_M\}$ and gene set $G = \{g_1, g_2, \ldots, g_N\}$

Output: Genes selected

Step1: Utilizing Relief feature selection algorithm to sequence genes, and generate candidate feature subsets;

Step2: Initialize the number of clusters, k, where k is an integer greater than or equal to 2. The fuzziness of the membership is controlled by m which takes value equal to 2;

Step3: Calculate the cluster centers of the FCM by Algorithm 2, the number of cluster centers is K;

Step4: Calculate the objective function min $J_m(U, W)$ according to Eqs.9;

Step5: If $||J_M^K - J_M^{K-1}|| \leq \varepsilon$, then go to step6, else go to step2;

Step6: For any attribute of each cluster, compute the correlation between A_i and A_j, $i, j \in \{1, 2, \ldots, p\}$, $i \neq j$, A_i and A_j are two attributes within an attribute cluster $C = \{A_j | j = 1, 2, \ldots, p\}$;

Step7: Calculate the significant factor $F(A_i)$ of an attribute A_i within an attribute cluster $C = \{A_j | j = 1, 2, \ldots, p\}$;

Step 8: Computation of *core* for each attribute cluster. For each cluster C_r, $r \in \{1,2,\ldots,k\}$, we set $\eta(C) = F(A_i)$, if $F(A_i) \geq F(A_j)$ for all $A_i, A_j \in C_r$, $i \neq j$;

Step 9: Termination. Steps 6, 7 and 8 are repeated until the $\eta(C)$ for the clusters does not change.Alternatively, the algorithm also terminates when the prespecified number of iteration is reached.

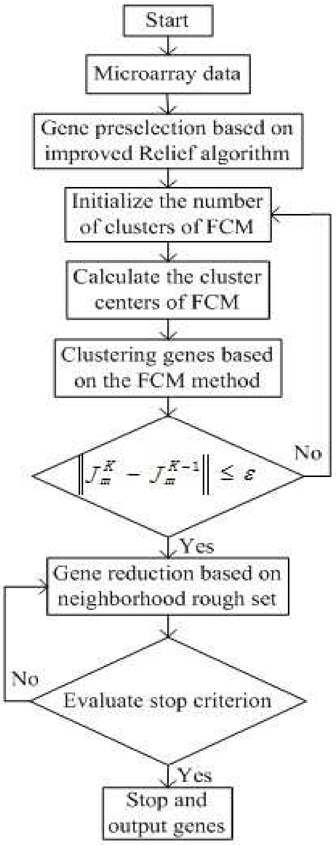

Start
Microarray data
Gene preselection based on improved Relief algorithm
Initialize the number of clusters of FCM
Calculate the cluster centers of FCM
Clustering genes based on the FCM method
$\left\| J_m^K - J_m^{K-1} \right\| \leq \varepsilon$
No
Yes
Gene reduction based on neighborhood rough set
No
Evaluate stop criterion
Yes
Stop and output genes

Fig. 1: The framework of the proposed gene selection algorithm (NMINR-FCM)

It is important to note that the number of clusters, k, is fed to the proposed algorithm as an input parameter. To find the best choice for k, we use the sum of the neighborhood mutual information measure $\sum_{r=1}^{k} \sum_{A_i \in C_r} NMI_\delta(A_i;\eta(C))$, to evaluate the overall performance of each clustering. With this measure, we can run the proposed algorithm for all $k \in \{2,\ldots,p\}$ and select the value k that maximizes the sum of the neighborhood mutual information correlation measure over all the clusters as the number of clusters. That is,

$$k = \arg \max_{k \in \{2,\ldots,p\}} \sum_{r=1}^{k} \sum_{A_i \in C_r} NMI_\delta(A_i;\eta(C))$$

4 Efficient gene selection algorithm

In this section, the performances of our proposed algorithm shall be demonstrated. In this experiment, the operating environment is Lenovo Windows7 PC with 3.1 GHZ CPU and 4GB RAM, and the algorithm is coded by VC++. In order to test the proposed algorithm, five different datasets which are from UCI datasets, used to study. A review of these datasets given in Table1 are as follows: (1) Dataset of breast cancer data is reported in [36], consists of 9,216 genes and 84 samples; (2) Leukemia1 is a collection of 7,129 genes and 72 samples, which is reported in [37]; (3) Leukemia 2 is another set of Leukemia [38], which contains 12,582 genes and 72 samples; (4) Small round blue cell tumors (SRBCT), reported in [39], are five different childhood tumors named so because of their similar appearance on routine histology, contains 2,308 genes and 88 samples; (5) Colon cancer contains 2000 genes and 62 samples, reported in [40].

Table 1: Gene expression data sets

Dataset	Genes	Classes	Samples
Breast	9,216	5	84
Leukemial	7,129	3	72
Leukemia2	12,582	3	72
SRBCT	2,308	5	88
Colon	2,000	2	62

In order to show the effectiveness of the proposed technique, several feature selection algorithms are compared. We conduct experiments with features selection algorithms of ReliefF [41], CFS [42], NRS [43] and NMI-EmRMR [33]. After features selection, three popular classification algorithms (Linear support vector machine (LSVM) and k-nearest-neighbor classifier (KNN) and CART) are employed for evaluating the quality of raw data and these selected genes.

In the following experiments, we will compare the numbers and classification accuracies of the genes selected with different algorithms. We normalize the feature values to the [0,1], and set $\delta = 0.15$ according to the experimental results. The statistics of classification performance is evaluated by 10-fold cross validation. Each dataset is first partitioned into 10 equal-sized sets. Then we use nine parts to create the training set for training the classification models and the remaining 10th to create the test set for evaluating the performance of each technique. In each training-test procedure of 10-fold cross validation, we repeat the algorithms used for comparison five times with different random seeds in order to ensure that the comparison among different classifiers does not happen by chance.

Table 2: Number of genes selected with different algorithms

Data	RAW	ReliefF	CFS	NRS
Breast	9,216	20	192	5
Leukemial	7,129	9	102	3
Leukemia2	12,582	15	150	3
SRBCT	2,308	7	70	2
Colon	2,000	6	46	2

Table 2 gives the numbers of genes selected with different algorithms, where raw denotes the feature numbers and accuracies of the raw datasets, ReliefF, CFS and NRS denote the results produced with ReliefF, CFS and neighborhood rough sets, respectively. From Table 2, we can see that only several genes are selected though gene selection algorithms. However, it is obviously that the number of genes selected decreases significantly. NRS just selects 5, 3, 3, 2, 2 genes for this task. The classification performances of these features are given in Tables 3, 4 and 5.

Table 3 gives linear support vector machine based classification accuracy computed with the raw data and the selected data, respectively. CART based classification performances of the raw data and selected data are shown in Table 4. KNN based classification accuracies, shown in Table 5, are similar with CART. From Tables 3, 4 and 5, we can see that ReliefF and CFS based gene selection algorithms are effective for LSVM and KNN based cancer recognition. However, they are not effective for CART. CART is much weaker in recognizing cancers than LSVM. Furthermore, NRS is not better than the other two gene selection algorithms. The results show that all the accuracies gotten from the raw data in Tables 4 and 5 are worse than those obtained with LSVM. However, classification performances improve much after gene selection.

Table 6 gives the number of genes selected and the performance based on NMI-EmRMR. Compared Table 6 with Table 2, we can see that the number of genes selected with NMI-EmRMR is more than NRS, and less than ReliefF and CFS. The results in Tables 3, 4, 5 and 6 show that the classification accuracies significantly rise with the genes selected with NMI-EmRMR. As to these gene selection algorithms, NMI-EmRMR is much better than ReliefF, CFS and NRS. Moreover, the genes selected with NMI-EmRMR are also more powerful than the raw data. The average accuracy rate rises from 90.3 to 93.1 as to LSVM, from 72.9 to 84.1 as to CART, and from 74.1 to 91.7 as to KNN. It shows that NMI-EmRMR is effective for gene selection.

The number of the selected genes and the corresponding classification performances based on the proposed algorithm NMINR-FCM are shown in Table 7. During the comparing between Tables 6 and 7, we can find that the proposed algorithm obtain better average accuracy than the NMI-EmRMR with similarity of genes selected. According to the experimental results presented in these tables, the proposed algorithm yields top-notch performance among these algorithms for all five datasets. For example, in Leukemia1 dataset, the accuracy value of the proposed algorithm is 98.8% as to LSVM, which is approximately 0.2% higher than that of NMI-EmRMR, 0.2% higher than that of ReliefF, 2.5% higher than that of CFS, 15.1% higher than that of NRS.

In these tables, average shows summarized result which is calculated by averaging the accuracy values over all datasets. The average classification performance of the proposed algorithm beats ReliefF by about 8.1%, CFS by about 3.5%, NRS by about 19.4%, and NMI-EmRMR by 2.1%, as to LSVM.

Table 3: LSVM accuracy of genes selected with genes selection algorithms (%)

Data	RAW	ReliefF	CFS	NRS
Breast	95.4 ± 8.4	84.5 ± 5.7	98.0 ± 2.1	67.5 ± 8.7
Leukemial	94.5 ± 6.5	98.6 ± 7.4	96.3 ± 5.3	83.7 ± 8.9
Leukemia2	94.6 ± 3.2	96.1 ± 6.9	95.6 ± 5.8	90.3 ± 6.8
SRBCT	82.4 ± 8.3	79.9 ± 8.3	86.0 ± 9.6	67.0 ± 7.9
Colon	84.6 ± 6.4	76.4 ± 9.8	82.6 ± 6.9	70.5 ± 5.3
Average	90.3	87.1	91.7	75.8

Table 4: CART accuracy of genes selected with genes selection algorithms (%)

Data	RAW	ReliefF	CFS	NRS
Breast	65.8 ± 4.7	76.3 ± 7.8	70.8 ± 7.5	77.5 ± 4.2
Leukemial	78.8 ± 2.6	93.6 ± 8.9	76.5 ± 6.3	88.5 ± 5.4
Leukemia2	90.3 ± 9.8	93.4 ± 8.6	88.5 ± 9.6	94.1 ± 9.8
SRBCT	65.7 ± 6.3	72.1 ± 3.5	74.2 ± 6.3	65.6 ± 9.0
Colon	63.9 ± 5.6	70.6 ± 7.4	74.0 ± 8.6	62.3 ± 6.7
Average	72.9	81.2	76.8	77.6

Table 5: KNN accuracy of genes selected with genes selection algorithms (%)

Data	RAW	ReliefF	CFS	NRS
Breast	68.7 ± 2.6	81.7 ± 3.6	97.5 ± 5.3	81.3 ± 6.2
Leukemial	82.8 ± 5.6	96.6 ± 6.5	97.5 ± 5.3	86.1 ± 2.7
Leukemia2	86.7 ± 2.3	94.3 ± 6.5	98.0 ± 4.3	93.2 ± 1.3
SRBCT	66.4 ± 9.8	79.0 ± 7.8	80.2 ± 9.8	66.5 ± 6.1
Colon	65.9 ± 6.2	75.9 ± 9.2	78.8 ± 8.9	69.9 ± 8.5
Average	74.1	85.5	90.4	79.4

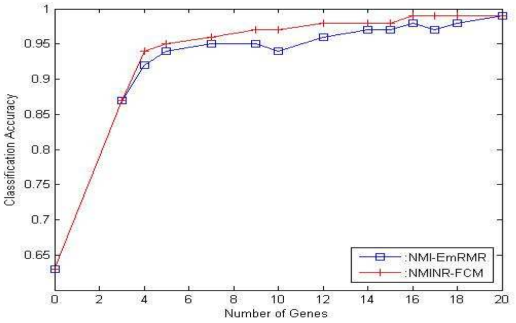
Table 6: Number of genes selected and performance based on NMI-EmRMR

Data	LSVM(%)		CART(%)		KNN(%)	
	n	Accuracy	n	Accuracy	n	Accuracy
Breast	18	100.0 ± 0.0	5	80.8 ± 10.4	15	98.8 ± 4.0
Leukemial	11	98.6 ± 4.5	2	94.3 ± 6.8	16	98.6 ± 4.5
Leukemia2	15	100.0 ± 0.0	17	96.5 ± 5.9	15	98.6 ± 4.5
SRBCT	9	84.0 ± 22.3	4	75.6 ± 3.7	14	82.3 ± 22.1
Colon	7	82.9 ± 5.3	2	73.3 ± 8.4	10	80.2 ± 9.2
Average	12	93.1	6	84.1	14	91.7

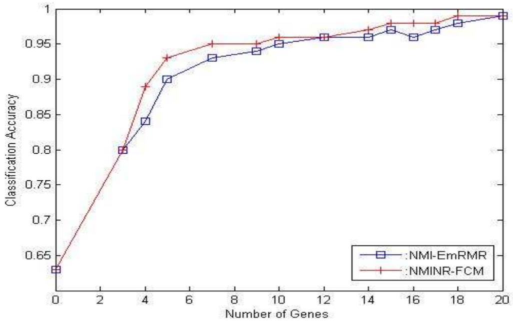
Table 7: Number of genes selected and performance based on NMINR-FCM

Data	LSVM(%)		CART(%)		KNN(%)	
	n	Accuracy	n	Accuracy	n	Accuracy
Breast	16	100.0 ± 0.0	4	84.7 ± 9.1	15	99.4 ± 3.5
Leukemial	10	98.8 ± 3.2	2	95.8 ± 8.6	12	99.2 ± 2.2
Leukemia2	12	99.6 ± 5.8	14	97.2 ± 7.8	13	98.6 ± 4.3
SRBCT	8	89.0 ± 9.6	3	83.2 ± 6.2	10	85.4 ± 11.8
Colon	6	88.6 ± 11.9	2	80.1 ± 5.3	5	84.9 ± 8.9
Average	10.4	95.2	5	88.2	11	93.5

Now we show the classification power of the first 20 genes selected with NMI-EmRMR and NMINR-FCM in Figs.2 and 3, where Leukemia1 and SRBCT are used as examples. Observing these two figures, we can observe that the NMINR-FCM is better than NMI-EmRMR. These results show that the genes selected can obtain high and stable classification performance. From the experimental results above, we conclude that our proposed approach is superior to other methods.

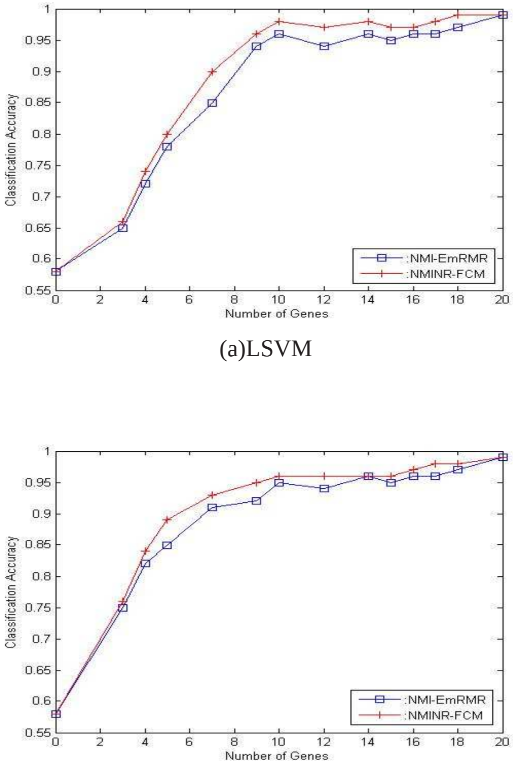


(a)LSVM



(b)KNN

Fig. 2: Variation of classification accuracy with number of genes (Leukemia1)



(a)LSVM

(b)KNN

Fig. 3: Variation of classification accuracy with number of genes (SRBCT)

5 Conclusions

Since gene expression data sets have thousands of genes and only a small number of samples, feature selection is an essential step to perform cancer classification, which to predict classes and a relatively small number of samples. Clustering and classification are key tasks of gene identification. Rough set theory has been used widely as an attribute selection approach. While the gene expression data sets are always continuous, the classical rough set methods cannot handle this case directly. Neighborhood rough set is introduced to deal with the continuous data in gene expression. In this paper, we introduce NMI to compute the relevance between genes and define the cohesion degree of the neighborhood of an object and coupling degree between neighborhoods of objects which based on neighborhood mutual information. Furthermore, the new initialization method of cluster centers for the Fuzzy C-means algorithm and the novel algorithm for gene selection based on Fuzzy C-means algorithm and neighborhood rough set are proposed. Five cancer data sets are gathered to test the proposed gene selection algorithm. Compared with ReliefF, CFS, NRS and NMI-EmRMR, NMINR-FCM gets good genes for cancer classification. Experimental results show that this algorithm outperforms than other approaches. In summary, we can get the conclusion that NMINR-FCM is effective and efficient for gene selection. In addition, we also find that the genes ranking the first several hundred are enough for cancer recognition.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Nos. 61370169, 60873104), the Key Project of Science and Technology Department of Henan Province (Nos. 112102210194, 142102210056), the Science and Technology Research Key Project of Educational Department of Henan Province (Nos. 12A520027, 13A520529), the Key Project of Science and Technology of Xinxiang Government (No. ZG13004), and the Education Fund for Youth Key Teachers of Henan Normal University.

References

[1] J.C. Xu, L. Sun, Y.P. Gao, T.H. Xu, An ensemble feature selection technique for cancer recognition, Bio-Medical Materials and Engineering, **24**, 1001-1008 (2014).

[2] S.Y. Kim, J.W. Lee, J.S. Bae, Effect of data normalization on fuzzy clustering of DNA microarray data, BMC Bioinformatics, **7**, 134-148 (2006).

[3] L. Sun, J.C. Xu, Feature selection using mutual information based uncertainty measures for tumor classification, Bio-Medical Materials and Engineering, **24**, 763-770 (2014).

[4] J.J. Dai, L. Lieu, D. Rocke, Dimension reduction for classification with gene expression microarray data, Statistical Applications in Genetics and Molecular Biology, **5**, 1-21 (2006).

[5] L. Sun, J.C. Xu, A granular computing approach to gene selection, Bio-Medical Materials and Engineering, **24**, 1307-1314 (2014).

[6] J.C. Xu, T.H. Xu, L. Sun, J.Y. Ren, An Improved Correlation Measure-based SOM Clustering Algorithm for Gene Selection, Journal of Software, **8**, 3082-3087 (2013).

[7] Z.X. Zhu, Y.S. Ong and M. Dash, Wrapper-filter feature selection algorithm using a memetic framework, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, **37**, 70-76 (2007).

[8] H.Y. Lin, Feature selection based on cluster and variability analyses for ordinal multi-class classification problems, Knowledge-Based Systems, **37**, 94-104 (2013).

[9] L. Sun, J.C. Xu, T. Yun, Feature selection using rough entropy-based uncertainty measures in incomplete decision systems, Knowledge-Based Systems, **36**, 206-216 (2012).

[10] M.A. Hall, Correlation-based feature subset selection for machine learning, Department of Computer Science, University of Waikato, Haimlton, New Zealand, (1999).

[11] J.C. Xu, L. Sun, Knowledge entropy and feature selection in incomplete decision systems, Applied Mathematics & Information Sciences, **7**, 829-837 (2013).

[12] H. Liu, L. Yu, Toward integrating feature selection algorithm for classification and clustering, IEEE Transactions on Knowledge and Data Engineering, **17**, 491-502 (2005).

[13] X. Li, H.S. Wong, S. Wu, A fuzzy minimax clustering model and its applications, Information Sciences, **186**, 114-125 (2012).

[14] R.J. Hathway, J.C. Bezdek, Optimization of clustering criteria by reformulation, IEEE transactions on Fuzzy Systems, **3**, 241-245 (1995).

[15] P. Tamayo, D. Slonim, et al, Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation, Proc Natl Acad Sci, **96**, 2907-2912 (1999).

[16] P.T. Spellman, G. Sherlock, et al, Comprehensive identification of cell cycle-regulated genes of the yest Saccharomyces cerevisiae by microarray hydridization, Mol Biol Cell, **9**, 3273-3279 (1998).

[17] J.C. Xu, Y.P. Gao, S.Q. Li, L. Sun, T.H. Xu, J.Y. Ren, A greedy correlation measure based attribute clustering algorithm for gene selection, Journal of Computers, **8**, 951-959 (2013).

[18] I.B. Aydilek, A. Arslan, A hybrid method for imputation of missing values using optimized fuzzy *c*-means with support vector regression and a genetic algorithm, Information Sciences, **233**, 25-35 (2013).

[19] X. Li, H.S. Wong, S. Wu, A fuzzy minimax clustering model and its applications, Information Sciences, **186**, 114-125 (2012).

[20] M. Xu, C.S Xu, et al, Hierarchical affective content analysis in arousal and valence dimensions, Signal Processing, **93**, 2140-2150 (2013).

[21] F.Y. Cao, J.Y. Liang, G. Jiang, An initialization method for the *K*-means algorithm using neighborhood model, Computers and Mathematics with Applications, **58**, 474-483 (2009).

[22] N. Belacel, M. Cuperlovie-Culf, M. Laflamme, Fuzzy c-means and VNS methods for clustering genes frommicroarray data, Bioinformatics, **26**, 1690-1701 (2004).

[23] J.F. Brendan, D. Delbert, Clustering by passing messages between data points, Science, **315**, 972-976 (2007).

[24] Z. Pawlak, Rough sets, International Journal of Computer and Information Sciences, **11**, 341-356 (1982).

[25] S.W. Zhang, D.S. Huang, S.L. Wang, A method of tumor classification based on wavelet packet transforms and neighborhood rough set, Computers in Biology and Medicine, **40**, 430-437 (2010).

[26] P. Zhu, Covering rough sets based on neighborhoods: An approach without using neighborhoods, International Journal of Approximate Reasoning, **52**, 461-472 (2011).

[27] Q.H. Hu, D.R. Yu, Z.X. Xie, Numerical Attribute Reduction Based on Neighborhood Granulation and Rough Approximation, Journal of Software, **3**, 640-649 (2008).

[28] S.L. Wang, X.L. Li, S.W. Zhang, J. Gui, D.S. Huang, Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction, Computers in Biology and Medicine, **40**, 179-189 (2010).

[29] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, Journal of Cybernetics, **3**, 32-57 (1973).

[30] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, New York, Plenum Press, (1981).

[31] N.R. Pal, J.C. Bezdek, On cluster validity for the fuzzy C-means model, IEEE Transactions on Fuzzy systems, **3**, 370-379 (1995).

[32] C. Budayan, I. Dikmen, M.T. Birgonul, Comparing the performance of traditional cluster analysis, self-organizing maps and fuzzy C-means method for strategic grouping, Expert Systems with Application, **36**, 11772-11781 (2009).

[33] Q.H. Hu, W. Pan, et al, An efficient gene selection technique for cancer recognition based on neighborhood mutual information, Int. J. Mach. Learn. & Cyber, **1**, 63-74 (2010).

[34] J. Li, H. Su, et al, Optimal search-based gene subset selection for gene array cancer classification, IEEE Trans Inform Technol Biomed, **11**, 398-405 (2007).

[35] Y.X. Li, J.G. Li, X.G. Ruan, Study of information gene selection for tissue classification based on tumor gene expression profiles, Chinese Journal of Computers, **29**, 324-330 (2006).

[36] C.M. Perou, T. Sorlie, et al, Molecular portraits of human breast tumours, Nature, **406**, 747-752 (2000).

[37] T. Golub, et al, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science, **286**, 531-537 (1999).

[38] S.A. Armstrong, et al, MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia, Nat Genet, **30**, 41-47 (2000).

[39] J. Khan, J.S. Weil, et al, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nat Med, **7**, 673–679 (2001).

[40] V. Alon, N. Rarkai, et al, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide array, Proc Natl Acad Sci USA, **96**, 6745-6750 (1999).

[41] M. Robnik-sikonja, I. Kononenko, Theoretical and empirical analysis of ReliefF and RReliefF, Mach Learn, **53**, 23-69 (2003).

[42] M.A. Hall, Correlation-based feature selection for discrete and numeric class machine learning, In: Proceedings of 17th international conference machine learning, 359-366 (2000).

[43] Q.H. Hu, D.R. Yu, J.F. Liu, C. Wu, Neighborhood rough set based heterogeneous feature subset selection, Inf Sci, **178**, 3577-3594 (2008).

Jiucheng Xu is currently a Professor at College of Computer & Information Engineering, Henan Normal University. He received his B.S. degree in Mathematics, Henan Normal University in 1986, the M.S. degree and the Ph.D. degree in Computer Science and Technology, Xi’an Jiaotong University in 1995 and 2004, respectively. His main research interests include granular computing, rough set, data mining, and intelligent information processing.

Tianhe Xu is currently a postgraduate in the College of Computer & Information Engineering, Henan Normal University, She received her B.S. degree in Educational Technology, Luoyang Normal University in 2011. Her main research interests include granular computing, data mining, and bioinformatics.

Lin Sun works at College of Computer & Information Engineering, Henan Normal University. He is currently a Ph.D. Candidate in School of Electronic Information and Control Engineering, Beijing University of Technology. He received his B.S. and M.S. degree in Computer Science and Technology, Henan Normal University in 2003 and 2007, respectively. His main research interests include granular computing, rough set, and data mining.

Jinyu Ren is currently a postgraduate in College of Computer & Information Engineering, Henan Normal University. She received her B.S. degree in Educational Mathematics, Shijiazhuang University in 2011. Her main research interests include granular computing, cloud model, image processing.